(4)

Some Brief Essays on Mind

Don Perlis

Technical Report 302
July 1989

UNIVERSITY OF
ROCHESTER
COMPUTER SCIENCE

SOME BRIEF ESSAYS ON MIND

Don Perlis

Technical Report 302
July 1989

This work was supported in part by the Defense Advanced Research Projects Agency (Office of Naval Research) under Grant N00014-82-K-0193, and in part by the National Science Foundation under Coordinated Experimental Research Grant CCR-8320136.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 302	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Some Brief Essays on Mind	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Don Perlis	8. CONTRACT OR GRANT NUMBER(s): N00014-82-K-0193
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Computer Science Department, Computer Studies Bldg. 734, University of Rochester, Rochester, NY 14627	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: D. Adv. Res. Proj. Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: July 1989
	13. NUMBER OF PAGES: 25
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Res., Information Systems, Arlington, VA 22217	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

none

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

mind-body problem, intentionality, default reasoning

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

These essays were written largely as notes to myself while I was on a very productive sabbatical leave at the University of Rochester, during the 1988-89 academic year. I am happy to asknowledge my benefit from discussions with Keri Jackson, Rosalie Hall, Alice Kyburg, Dorit Bar-On, and Elizabeth Hinkelman; these discussions prompted a great deal of what I have to say here. Still, the essays remain in highly preliminary form, not yet worked out to the extent that I would like, and certainly not to be taken as representative of the views of Keri, Rosalie, Alice, Dorit, and Elizabeth.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (Continued)

The essays are presented here in the order in which they were written: 1. Robots, Us, and the Future; 2. Pains and Brains; 3. Brain, Knowledge, and Experience; 4. Intentionality and Defaults.

SOME BRIEF ESSAYS ON MIND

Don Perlis
Computer Science Department
University of Maryland
College Park, MD 20742

These essays were written largely as notes to myself while I was on a very productive sabbatical leave at the University of Rochester, during the 1988-89 academic year. I am happy to acknowledge my benefit from discussions with Keri Jackson, Rosalie Hall, Alice Kyburg, Dorit Bar-On, and Elizabeth Hinkelman; these discussions prompted a great deal of what I have to say here. Still, the essays remain in highly preliminary form, not yet worked out to the extent that I would like, and certainly not to be taken as representative of the views of Keri, Rosalie, Alice, Dorit, and Elizabeth.

The essays are presented here in the order in which they were written:

1. Robots, Us, and the Future

2. Pains and Brains

3. Brain, Knowledge, and Experience

4. Intentionality and Defaults

Robots, Us, and the Future

I wish to try to explain my view of artificial intelligence, and more broadly how it fits into science as a whole. In doing so, I will not hesitate to indulge in sheer speculation when it seems to fit the topic.

I will begin with a negative thought (one that I do not agree with). Consider the statement that, while robots and AI (artificial intelligence) may make great strides in the future, still they never will be able to produce music with the sensitivity of certain humans with great musical talent.

First, I am reminded of bold statements made in the past: people will never fly, will never understand the incredible chemistry of life, will never go into space, will never run a 4-minute mile... However, what I want to argue is a rather different position, namely, that it is not so much a matter of replicating ourselves or building devices to take over the more unpleasant of our tasks (worthy a goal as that may be) as it is of *understanding* ourselves and in so doing changing our very concept of ourselves. We may get to the point of knowing enough about ourselves to be able to make major changes in what we are, we may control our own future, our evolution. We then will be machines, in that we will construct ourselves. We already do this in small ways today, e.g., by body-building, by education, by medicine. But we may someday be able to choose much more profound aspects of our being, and, even more exciting, we may thereby come to see entire ways of being that today totally elude us! We may find that we are merely a rather minor form of consciousness in a vast spectrum of possibilities waiting for us to grow into them. This may even lead to an understanding of understanding, so that perhaps new and deeper and richer forms of understanding become possible, combining imagination and feeling and intellect or even other modes we cannot yet dream of.

The dull robots of today's factories are a very poor image for this prospect, and it invites confusion to use the same words (robot, computer, machine) for these as for the future scenario of ourselves I am suggesting. But in another sense the words are appropriate, just as chemistry is an

appropriate term for describing the basis for life. It is chemistry, but a rather fantastic version of it not suspected until rather recently, one that puts 19th century chemistry to shame as a dull dead affair. In the same sense I think that the possibilities for machines will turn out to be even more dramatic and we will see that we are not only chemical in nature but also mechanical -- yet this will not detract from our wonderfulness at all, any more than chemistry has. We will find that we are wonderful soft machines, and this will attest to the amazing qualities machines can have: they can be as wonderful as us: they *are* us! It is just that so far we have not come to realize the incredible range of things machines can be, we have at present only rather pitiful examples, just as people of the past had only pitiful examples of what chemistry could be.

To return to our initial theme: Skilled musicianship is most likely a fairly recent phenomenon; e.g., quite possibly 10,000 years ago there was little if any music worth the name. And so it is with most things we regard as important elements of our lives. But then how likely is it that we now have reached pretty close to the limits of what is important and worthwhile, stirring, central, etc, about living? Isn't it more likely that we have merely scratched the surface of what life can be, that our sensitivities themselves are primitive compared to what they can/will be 10,000 years from now? And consider that technological developments are part and parcel of increased sensitivity, better acoustics, precision instruments that simply could not have been made 10,000 years ago, etc.

So it is with AI. By better understanding the basis of intelligence, thought, mentality, we ourselves (our children, our creations, our choices) evolve into fancier instruments, more complex. sensitive, rich, varied. The realm of possible modes of being, of feeling, of communicating, are waiting to be discovered, and AI is the science of discovering them. It is not at all a matter of being 'replaced by machines' -- we are machines ('meat machines' in Minsky's phrase) and it is a matter of understanding in detail just how we work, how it is we think and feel, that will allow us to grow beyond our current state. For instance, we may learn (perhaps for future generations to use) how to greatly extend our auditory sensitivity, or our ability to communicate with one another -- perhaps

even to the point of being able to feel one another's thoughts and feelings!

Some may regard this as dangerous, for us to have that much power over our future. But it also is a tremendously exciting prospect, to think we may conquer what it is in us that leads to war, poverty, unhappiness, loneliness. Wouldn't it be wonderful to be able to control these? And even more basic, our very natures, that seem so solid and unquestionable, may open up into realms we cannot even dream of today. Moreover, we may find out what feelings really are, what happiness is, what togetherness and wholeness (as feelings, as mechanisms, as parts of being) really amount to. We may discover morality in terms of higher principles of consciousness, rather then than hit-or-miss individual guesswork of today. But this surely cannot happen without growth on our part, enormous growth far beyond our current limited imaginations. We may need entirely new concepts, far more subtle things than our current phrases of 'conscious' or 'good' etc. This is similar to the development of modern biology in which the concept of 'life' has given way to a far more subtle range of concepts, involving complex matters of self-reproducing molecules and multi-cellular development rather than a single all-inclusive term. This is not to say that we have discovered all there is to life, far from it. In particular, the entire cognitive and emotional areas are pretty much mysteries at present. But this is where AI (and psychology) are aimed.

Of course, new developments bring risks, but so does staying put. Avoiding new developments 10,000 years ago would mean no music today, and no poetry, and no science, and no medicine. Surely we are not now at the perfect spot so that we should stop. And we would be stopping at the brink of the answering one of the greatest questions of all: what are we?

One part of this general issue is the so-called problem of intentionality. Although it falls short of the problem of 'feelings' I think it is a related matter: what is it for an agent (machine, person) to intend something by an action, and in particular, what is it to intend a meaning by a word? That is, how is it that when we say 'my dog' we *mean* an actual dog, whereas an ordinary (!) program with those words wouldn't mean anything by them?

It seems that *our* programs (our brain processes, in our heads) have features (intentionality) that, so far, our written 'computer' programs lack. And it is this kind of computer program 'dumbness', I think, that makes people tend to think that programs and machines can never have our sense of identity, our kind of integrated, vivid, personal subjective experience, never 'really' think, feel, believe, hope, etc. But I think that we simply have not yet learned how to build very complex kinds of programs, largely because we (and psychologists) have not figured out how *we* work! So AI is also a hand-maiden to psychology, in providing tools for building complex models that would be incredibly time-consuming and tedious without computers. However, I think that the intentionality aspect of the more general problem of mind has at least partly been solved.

Thus I see AI not as trying to replace or outdo us humans, nor as trying to make our lives easier by having robots do menial tasks for us, but as a way to understand ourselves and thereby change/grow, just as we always have been doing -- although each new phase brings revolutions in how we see ourselves and in fact in what we are physically, psychologically, biologically.

Even consciousness itself is a mystery, and once we understand it I think we will have reached a new age of being. We may look back then on the present as a very primitive state of being, where we each seem to be a separate entity standing out against the world as backdrop for our personal ambitions, rather than us being those parts of the universe that have marvellously come conscious through eons of evolution so that the universe can finally look at itself (with us its eyes, so to speak). I think some of the religious ideas veer a little in this direction, about brotherhood, oneness, wholeness, being, morality, meaning, etc. But these usually are presented as cut-and-dried, as already understood and to-be-accepted, rather than full of mystery and a matter for investigation. We all recognize that existence is a mystery; but what we do not have is an answer to the mystery. The *quest* is in effect a religious one, though not in the customary sense of being tied to a particular doctrine, since we simply do not yet know very much on which to base firm beliefs.

However, this need not be left to theologians: We Alers (and scientists in general) are positioned to do something significant with it, to discover a new age of awareness by finding out what being human (conscious, feeling, etc) really amounts to! In fact, many people have turned to science precisely because they wished answers to fundamental questions about the meaning of life, what we are, what it all amounts to. I think our li' es here on Earth have little if anything to do with broad meanings of the existence and the universe as a whole, at least *so far*. But if we can vastly expand our knowledge, our understanding, our range of concepts, values, natures, perhaps someday we will be in a position to understand such questions, and even to answer them! Progress is slow, but I think it's better than telling ourselves that we already know the answers, when we are currently highly ignorant of the broad features of what is what, ignorant even of a range of concepts appropriate to understanding those features.[1]

Science has hardly begun to address the idea of what counts as an explanation; that is, we do not yet have a scientific grasp of what explanations are. Science uses (makes) explanations, but explanations themselves are not the subject of current science. Once they are, we may have a very different set of conceptions as to what the question 'what does it all mean?' means. A sample of what I am driving at can be given by considering the idea that, whenever we explain X by Y, we can then ask for an explanation of Y in turn, thus never reaching anything 'final'. Perhaps we will someday be able to study (infinite) hierarchies of explanations, and come tc an understanding that transcends this regress, reaching something final after all. Indeed, a few scientists and philosophers are already broaching this kind of thing. This is related to the idea of 'wholeness' that I mentioned earlier. This has in the past seemed mere speculative whimsy, but recently it has taken on a more precise character in developments in physics, where a purely part-like picture seems unable to account for empirical results.

---

[1]As an aside, it seems that this aspect of scientific investigation is the one so hard for some to accept: that we are truly ignorant and can still go forward, make progress toward reducing that ignorance. It is, to be sure, scary to face an unknown world, it is tempting to comfort oneself with easy answers; comforting and tempting but little more than that. Yet science has, slowly, shown us glimpses of marvels yet to come, and that itself is sign for hope and enthusiasm to many.

A related point: It is sometimes said that science has nothing to say about values or ultimate meanings. But I disagree. First of all, as I just said, many scientists are motivated to do science by concern for values, to understand our being, including what it all amounts to. This surely is a value of a very high order, a quest for ultimate truth. However, this only indicates that science may be a value in itself, not that it is about values. Still, as I have emphasized earlier, science (especially cognitive science such as AI and psychology) may someday be able to show us what values really amount to: feelings, morality, meanings, understandings, and so on. And in so doing, it may change our picture of being so much that our current questions about values (which values are good, what is the ethical life, what should I do) may take on a new, profounder, better understood quality from which we may then see real answers (perhaps to better, profounder questions). The fact that science has not yet succeeded (indeed barely begun) in the effort to understand values, is no argument that it will not do so in the future.

One aspect of this is the study of complex systems, something we have little experience with so far. Few would contest that the mind is a highly complex system. Yet we know little of how complex systems work; from a description of the (workings of the) parts an understanding of the whole is not at all immediate. (Indeed, as I mentioned earlier, understanding itself is something that so far is just a folk-concept, that may give way to a profound range of awarenesses, perhaps even breaking down the traditional division between thought and feeling.) It may be fair, then, to characterize AI as a branch of the general science of complex systems. Yet complex systems, as we are discovering in biology and now in so-called *chaotic* systems, are far richer than we had previously suspected. The mind may well prove to be the most complex (and correspondingly rich) of all, a system in which values play a central role in a way we do not yet comprehend.

To encapsulate: All our concepts may alter profoundly as we find out more about time, space, cosmology, life, mind, and how how it all fits together in a mysterious whole we currently have only the tiniest glimmer of. We are parts, everything we usually deal with is just a part, yet existence itself is a whole of which we are parts, though we have so far only a limited sense of

what wholeness amounts to. Moreover, much of the history of science has been toward ever-more-inclusive and general theories, even to theories of theories. Thus we are already underway (although just barely) on this fantastic effort to see what it ALL amounts to!

Pains and Brains

At one point in his argument against the mind-brain identity thesis (in Naming and Necessity) Kripke makes an observation central to his position. Namely, he says, it is very implausible to think that the mere firing of some brain fibers is, in itself, the phenomenon of feeling pain.[2] He is careful to insist that pain is by definition something felt, something going on in a feeler (e.g., a person). And indeed it does seem implausible that the mere firing of a few little fibers could then be felt-pain, for who then is feeling it? And if someone (other than the fibers themselves) is doing the feeling, then it is not simply the firing fibers in themselves that constitute the pain. For pain, unlike his example of heat, is by its very nature a feeling. We should perhaps say, rather than that pain is felt, that the agent is in a state of pain. It takes an agent to be in a state of pain, then, and no mere fibers apart from an agent can constitute this.

Now I think this is a pretty good argument. But Kripke takes it further, to say that therefore fiber firing cannot be pain. There is a tiny but crucial slip here. For we must first determine that these fibers are indeed 'a few', 'little', and 'apart from an agent' as stated above. There is an underlying assumption, I suppose, that fibers are microscopic things, insignificantly modest portions of the whole brain, so that even if brain process as a whole were to constitute mind, surely no mere fiber or two could.[3]

But we then can make a valid and sound inference, I think, to the effect that this cannot be a very small part of the brain. That is, no very small part can, when suitably activated, constitute pain, for it would then also have to constitute a person, in the sense of a feeler, one who feels the pain: it would have to be a kind of mind. And surely the mind, if it is a brain process, is not a sim-

---

[2]Kripke actually speaks of C-fibers. In fact C-fibers (and A-delta-fibers) are carriers of 'pain signals' from many portions of the body to the spinal cord; they are not in the brain at all. No anatomist seriously thinks that firing of C-fibers is itself the feeling of pain. The spinal cord relays the signals from these fibers on to the brain, especially to the thalamus region, where -- some have speculated -- actual pain sensation may occur. We will therefore not follow what has become standard philosophical mis-nomenclature, but rather refer to 'pain-fibers' firing in the brain; whether there are such specialized pain-fibers at all is then in part what we are addressing.

[3]The fibers in question are neurons -- i.e., single cells -- and indeed small when taken one at a time. More on this below.

ple one confined to a microscopic bit of brain tissue.

Kripke's argument then seems to come down to the tacit claim that a few little ol' fibers surely are incapable of conscious feeling. They can fire without their (or anyone's) being conscious. But then they cannot constitute felt pain. If we take Kripke's notion of pain as felt pain, this makes his point: (a few) brain-fibers firing is not pain. (In fact, this is not even a necessity/possibility issue: the point really argues that firing a few microscopic fibers is not the feeling of pain, period.) Yet would Kripke so confidently say the same of the whole brain -- at least 100 billion neurons strong -- or indeed any large portion thereof? To do so is simply to beg the entire question. Maybe the whole brain does constitute a feeler, a conscious agent. Indeed that is what the identity thesis claims.

Thus Kripke's argument in no way undermines the identity thesis. Rather it gives us new information about what must follow from it: that mind (and feeling) is a global property of the brain, and highly local events in themselves cannot constitute conscious (cognitive or mental) events. In fact this is somewhat borne out by what I understand is shown in recent studies using tomography and other brain imaging techniques. Traces of activity levels in the normal living brain show that when a person listens, or thinks, or is pricked with a pin, very large areas of the brain change their activity level (oxygen or sugar consumption).

But what, you may ask, about the Penfield studies showing that stimulation of just a few neurons in one part of a person's brain can cause that person to hear a symphony, or see colors? This I think does not in any way refute the above claim, for we can simply say (what is very plausible) that stimulation of even one neuron can (and usually does) lead to the stimulation of millions (and even billions) more via the massive interconnections between neurons. Thus the original stimulation of one neuron is not in itself the hearing of a symphony, but rather is the event which triggers the massive stimulation of the brain as a whole which is the hearing of a symphony.

This is not to say that only the entire brain can constitute mind. Indeed rather large portions can be absent and yet consciousness remain intact (though perhaps diminished). Nor is this a claim

of centrality, that some central whole piece of the brain constitutes mind. For each hemisphere alone is capable of consciousness. But at the very least it is a claim that many millions of neurons must be intact and involved in any conscious state. Feelings are global properties of the mind itself. They are not external entities the mind looks out on from afar.

What then about the thalamus and pain-fibers in the brain? According to our analysis, the actual experience of pain -- felt pain -- can occur only as a state of a feeling being. Thus the thalamus in and of itself in a certain state cannot be what pain is unless the thalamus itself constitutes a feeling agent, something no one at this point seems prepared to say. However, the thalamus is constituted of billions of fibers (neurons), so that this is not out of the question.

(1) Kripke, S. [1980] *Naming and Necessity*. Harvard.
(2) Thompson, R. [1985] *The Brain*. Freeman.

Brain, Knowledge, and Experience

Thomas Nagel and Frank Jackson have argued that science does not provide us with certain kinds of information, namely knowledge of what subjective experience is like. Nagel argues, for instance, that all our detailed examinations of a bat's brain will still leave us in the dark as to what it is like to be a bat.

Does science really abstract away from individual subjective experience? This seems to be the case regarding issues of verification: science demands repeatability by many observers.[4] But in principle one sole scientist could go it alone, and do fine. I don't think Nagel's argument really rests on this at all. The point is rather than none of the observers (in the enterprise of science) is a bat (and if one were, that one presumably could not communicate all its findings with the others).

Now, none of this is to say that the subject matter of science does not include subjectivity itself. The topic of a bat's feelings (how it feels to be that bat at that moment) is quite different from the abstraction away from subjectivity regarding verification. Now suppose, as Nagel permits us to do, that we find out what bat feelings are, in ordinary scientific terms. Then we do not necessarily know what those feelings feel like. But I want to argue that we might, the long way around. It may be a little like sympathizing with an unusual character in a well-written novel, where the author has to go to some lengths to draw the appropriate picture for us. We of course have to be able to follow it, and this may tax our memory and attention and so on, to the point that we may need mechanical memory aids. It may even, in the case of bats, require us to use bat-oid memories temporarily, to be able to hold onto the large number (suppose) of sounds all at once.

This I claim is not cheating. The same holds for many (most?) things we understand or know. Cloud formation, for instance, is only understood by us in a vague and approximate way, but not because of any profound ontological features; rather simply because clouds are so complex as to

---

[4]How similar these observers must be is precisely the point. We must at least be able to agree on certain elementary data (inputs and outputs). We may also need sufficiently similar conceptual frameworks to be able to compare ideas. I think that this may be a key to the whole issue.

defy detailed analysis. So we use simplifying mathematical assumptions, and point to that instead. If we really wanted to grasp the process of cloud formation as it really is, we would need to enlarge our short-term memories enormously to account for each droplet of water vapor all at once.

A related illustration is visio-numerical apprehension. Most people can judge the number of a pile of up to four objects without counting. The visual system seems to do this automatically. But few of us can do it for 10 or more objects, though some can. What does it feel like to have the latter ability? We may someday understand in total detail just what the ability is, in terms of brain structure. But this will not necessarily let us (those who only reach to 3 or 4) have the feeling, unless we can train ourselves to do it too. That is, there are limits on any given brain's computing powers.

This seems to be in support of Nagel. In part, yes. I claim that a typical fact about feeling like something (a bat, or a 10-apprehender) is very complex, on the order of cloud formation, and too complex to grasp in its entirety without special purpose hardware. That is, understanding is itself constrained by the brain's hardware. We rely on approximate or fuzzy versions of complex things when we try to understand them. To feel like a bat one needs the relevant aspects of a bat's hardware. To know what it is like to feel like a bat is simply to feel like a bat (at least in part, in imagination), and to feel like a bat is to have batty feelings in one's repertoire of feelings, i.e., to be able exhibit those relevant aspects. Similarly, to see one has to have a visual cortex (or suitable surrogate). To bat-tize one has to have bat-cortex or surrogate.

But this is not to say that science cannot give us this knowledge (this feeling). It can, if we are prepared to use crutches to aid our brains. Implicit in the Nagel-Jackson view seems to be the idea that there is a clearcut and complete sort of pristine rationality that, given enough time and space, can exhaust all that is worthy of being called science. This sounds like a Baconian view of science: the compiling of a master catalog of local details so that all patterns can be seen. But to see a pattern often takes special ways of looking. The mere sequence of integers does not in itself single out primes: someone has to 'see' the prime property. Similarly for clouds: mere specification

of atomic locations does not amount to a grasp of cloudhood. Special apprehensional ability is needed for virtually all scientific (or other) knowledge, whether by virtue of insight (as in the case of primes) or of hardware alone (as in the case of bat feelings).

To know what it is like to see the world through rose-colored glasses, one can put on a pair. Does one now know a new fact? One now has a new datum, the rose-world-appearance. One knows the datum, in the sense of knows-of, is-acquainted-with. It is like knowing the number 2. One knows of it, one doesn't know that it. The bat-feeling is a datum. Certain data require special equipment to capture. Indeed, all data do. We often capture data indirectly with equipment, e.g., Geiger counters, since we lack built-in devices. So, Nagel can be seen as merely having shown that the unaided human brain cannot record all data. And when we use indirection, we do not ourselves acquire the datum. For that, our brains would have to develop effective and appropriate integration of those devices.[5] Given all the initial (local) data, one can calculate a lot of things. But can one calculate chaotic phenomena resulting from the initial data? Perhaps so, perhaps not. Until recently, certainly we humans could not; it took insight to see the existence of chaotic phenomena. So possibly a brainier species than us but with the same local facts as us, may well see consequences of these facts that elude us. Yet they are using rationality too, the difference is that they may have a richer rational format. It is not at all obvious that there is a complete, pristine rationality that reveals all 'physical' facts, unless one defines physical so narrowly that it leaves out most of modern science.

In principle, our brains supposedly can compute anything (that is computable at all: the Church-Turing Thesis). But this is a worthless observation as far as providing a scientific tool, for we would have no idea which of the infinite succession of computations we could perform would actually be worth performing. We would not know which of them had anything to do with bats, or primes, or clouds. We need insight to guide our attention to key computations. We might stumble

[5]Of course, whether we would be able to remember the bat-feeling if we return to 'normal' is unclear, for it would depend on whether the memories can be stored in normal human brains, and this may well be false. But this in no way bears on science; human brains develop too, and every time we learn something our brains change a little.

on things by luck, of course. But the more complex the phenomenon, the less likely the lucky stumble; and in certain cases, as already mentioned, our brains may be too small to record such a computation at all.

But it is still possible that the Nagel-Jackson thesis is right: maybe even all the computing power in the world cannot compute a bat-feeling. Maybe qualia are beyond computation/rationality of any sort. But consider what that would mean: that even if we managed to get our brains to perform the computations (physical processes) in a bat's brain we would not have the bat experience. This seems to be strongly veering into substance dualism, far from the property dualism Nagel and Jackson espouse. It seems to say that it is not the mere physical behavior of the bat brain on which bat feelings supervene.

We may have been taking terms like 'understand' and 'know' for granted, as if we had the idealized pristine rational engine all sewed up for good. But we typically form highly simplified models when we 'understand' things. If we want to understand in a fuller sense, it is not science that says no but simply our limited brains. The fact as to what physically constitutes a bat-feeling is probably utterly enormous in complexity. If we were able to grasp it all at once, as in a 10-apprehender, we then might very well be, for all practical purposes, a temporary sort of bat. But grasping something all at once is not a matter of having a new fact, except in the rarified sense of a compound or complex fact. The fact of the entire bat-feeling may involve so many parts that we can at best catalog them, and call them to mind one by one. The bat brain does them all at once, however: like the cloud, it *is* them! So the compound fact is not necessarily known by us all at once. Some complex facts (indeed, quite possibly most) may be beyond our currently-evolved brain capacity.

Consider an example due to Frank Jackson: Mary lives in a black-&-white room and sees no colors but has access to complete scientific knowledge, including knowledge about wavelengths of light and about people's eyes and brains outside her room, and can figure out all facts (so the story goes) expressible in the language of science. Jackson claims that Mary will not know what it is like

to see color.

Yet I claim that all we are entitled to conclude is that poor Mary, smart as she is, simply will not know every *compound* fact all at once, unless she has evolved an infinite brain capacity. But if she did, or even if it were just very, very much larger than normal, she then would be able to know what it is like to see color, be a bat, etc. She would merely need to take the pains to assemble all the right stuff in her memory in the right way, as dictated by her catalogued understanding of these things. That would constitute her actually seeing and feeling these things just as people and bats do. And it would be understanding; experience *is* understanding, but with hardware of large grasp, so to speak. Our folk-psychologic terms 'understand' and 'experience' are too sloppy as yet. When we learn to characterize them better I think we will find that there is no primitive bottom-line level of understanding that science necessarily rests on. It is simply whatever scientists' brains do readily, and thus is a kind of experience. Similarly, experience is a kind of understanding. Thus, for instance, seeing red -- knowing or having the experience of red -- may involve millions of neurons firing all at once in a special pattern. Mary would have to account for this complex pattern in detail in order to 'know' the experience. Either her visual cortex does it for her, in the ordinary way, or she has to work it out on the basis of 'science'; but in the latter case a mere high-level summary may not be enough, she may need to get the details right. And as with cloud formation, the details may be too much for an ordinary human brain.

If physicalism[6] is correct, then the kind of small grasp that our brains have is sufficient to *catalog* all (local) facts, even if not to grasp all compound (global) facts. Our poor grasp of grasp leaves us today in the clutches of a seeming dualism between objective (understanding) and subjective (experiencing). This is what Nagel and Jackson have unwittingly exploited. The weakness of their arguments will be seen as our folk psychology matures. Of course, this may require us to grow larger brains!

---

[6]I.e., reductive physicalism, in which truth is ultimately a catalog of local details (except for spacetime relations among these local details). Yet Bohm, Buchler, and Wheeler -- among others -- have argued that physics and philosophy must embrace an idea of wholeness (globality) if they are to explain the full range of experience. Bell's Theorem is thought by some to provide a formal (and empirically verified) argument for globality.

Scientists rely essentially on subjectivity, that is, on experience. They need to be able to single things out, to pick out a datum from the background, they need attentional memory. No amount of 'abstracting away' can get around this. If we take the notion of 'fact' as something relative to a thinker's kind of subjective grasp, then we are led to a Nagelian (and Putnamian) view. But physicalism (and realism) assert (roughly) that there is an outer factual reality. So far, it seems that this can hold up. Interestingly, however, the idea of compound fact is also essential, for we need to be able to take cloud formation, and bat feelings, to be large-scale collections of facts, even if we do not grasp their entirety. Thus science depends on a large-scale language, such as set theory, to discuss wave motion, composition of stars, electricity, etc. The 'togetherness' of facts is also a fact, and a very important one, without which there would be no clouds, no bats, no scientists.

(1) Jackson, F. [1982] Epiphenomenal qualia. *Philosophical Quarterly*, 32, 127-136.
(2) Nagel, T. [1974] What is it like to be a bat? *Philosophical Review*, 83, 435-50.

Intentionality and Defaults

Mind is a device for reasoning, thinking. So, what is thought? Who needs it? Not bacteria. But more complex behavior seems to require processing information 'about' the world. What is 'aboutness'? -- and what good is it?

The world is too complex to always correctly model it or algorithmize always appropriate responses to it. For bacteria it seems not to matter, they survive in sufficient numbers without having to deal with this issue. But we are not so lucky, or rather, we are lucky that we are not so lucky, since it has forced us to evolve ways to deal with incorrect algorithms, namely, to postulate error in ourselves and, on detecting it, take corrective action.

This is the essence of default reasoning. Thus -- to venture a strong statement -- defaults may be the reason we have minds. I will consider the extent to which aboutness may be explained in terms of this capacity, and apparent advantages it confers.

Commonsense reasoning is I think now widely recognized as largely coming under the general province of default (or defeasible) reasoning. Most attention on this has been aimed at capturing the use of defaults, i.e., of getting the right defeasible conclusion, and not on what to do when one finds a defeater for a previous conclusion. The latter is what I have called the 'fix' problem. A 'simple' solution is to throw away the defeated conclusion; but it is an unwise solution. A better one is to make defeat (and error) itself a topic about which one has extensive knowledge, for then one can explain how past errors occurred, learn to avoid similar ones in the future, and in general take account of distinctions between appearance and mention and belief on the one hand, and reality and use and truth on the other. The representation of error and the appearance/reality distinction lead directly into the above issue of aboutness. Thought is, to a large degree, a matter of distinguishing what is from what isn't, and in particular what is the case from what one merely *thinks* is the case, i.e., a matter of recognizing the possibility of error.

Standard formalisms for default reasoning do not provide conclusions about error: they either provide a default or they do not; they do not give meta-assessments of the state of one's reasoning.

Here are some examples in default reasoning where explicit representation of the appearance/reality distinction (ARD) is useful:

1. The Nixon Diamond (due to Ray Reiter): Nixon is both a Republican and a Quaker. Quakers typically are pacifists, and Republicans are not. Is Nixon then a pacifist or not? Based on the given information alone, an intuitively plausible outcome is not simply to believe nothing at all, but to realize that there are two possible outcomes (pacifist and non-pacifist) of interest and that we do not know enough to decide between them. Thus we recognize that there are two appearances (pacifist and non-pacifist) and only one of them can be true. Also we refrain from repeating the effort later -- we either leave it alone or seek more data.

2. We have a pile of seeds, which we are dropping one by one onto the ground. We release our grasp on a seed, expecting it to fall. But it sticks to our fingers. The belief that it would fall is seen to be untrue, and so we try again, this time using other means. But it may be important to remember that our first try was based on a false belief, if we now want to drop another seed. Should we proceed as we first did, or assume it too may stick? It may not matter, we can go through the whole thing again, except that this is rather slow and unintelligent behavior. It is especially vivid in the case of water on our fingers which if we notice it can lead to our drying our fingers with a towel. But if we dry our fingers and it turns out that the pile of seeds is wet then the towel has been useless. So why did we bother drying our fingers? We cannot tell, unless we remember our thinking.

3. We travel to Lower Slobbovia, and see various birds. The first one we see does not fly away when we approach quite close, contrary to our expectations. If we do not remember this, and yet continue having similar experiences, we will have no reason to alter our general expectations of birds in Lower Slobbovia. And if an unconscious weighting mechanism keeps count of defaults gone wrong (in a way that does not interact with our database and inferences) until a threshold is reached and then the default is removed altogether, then we will be in bad shape when we return

home from Lower Slobbovia. That is, it is important to note that a default is not working (and even to note the circumstances). Moreover, the next time we visit Lower Slobbovia, we will want to take a special lens for close-up pictures of birds. So we need to have the revised rule explicitly represented. Finally, we end up with excellent close-ups of Slobbovian birds from all 17 of our Lower Slobbovian vacations except the first. How come? We cannot explain this except with reference to our early mistaken expectations.

This might have important repercussions. Suppose we are being questioned in court as to why we purchased a close-up lens just before our second trip to Lower Slobbovia -- it is alleged that we intended to take photos of top-secret documents. We claim it was to photograph birds.[7] But then why did we not buy it before our first trip? Because we learned that Lower Slobbovian birds cannot fly during our first trip, not before. Consider what the court would think if we cannot recollect this, if we actually are puzzled ourselves since we have no recollection of having had a false belief about Lower Slobbovian birds. Thus our own interests are again served by retaining information about our own thinking activity.

Another, perhaps more practical, consequence can be seen. If we want to know, for insurance purposes, what year we bought the new lens, we can figure it out by remembering that we still believed that Lower Slobbovian birds could fly until we were already on our first trip, so the lens must have been bought after that. Of course, one could simply remember perfectly when one buys things, and not need such fancy reasoning. But if we postulate a perfect memory, then surely it is odd not to allow memory of one's course of reasoning as well.

These examples illustrate that making long-term use of experiences gained in novel situations, is enhanced by having high-level access to the course of those experiences, including false starts and other errors. Still other examples can be suggested -- some of which are already in use -- such as real-time issues of taking account of where one currently is with respect to a task, which seems to hinge in part on the ARD in order to separate one's goals (which are in the realm of

[7]To know even this about ourselves already requires storing information about our past mental activity.

beliefs or thoughts) from one's current state of progress (the reality).

Also, the problem-solving technique of experimenting to see what works seems to crucially involve this same feature, especially when using a judicious mix of random trials and thought-out prospects. It makes little sense to rely purely on random trials, yet pure advance-planning is often very slow as well. A mix seems to come closer to what people do, planning a general range of likely possibilities within which to experiment, and also letting the results of the experiments re-align one's assessment of future likely possibilities. It's much like best-first search, except that it may be directly coupled to action in the environment. And then marking an experiment as evidence that something did or did not work can be very useful for future reference.

A very different realm in which ARD reasoners should excel is natural language processing. Assessing differences in usage between speakers is a canonical case of the ARD, for a word must be distinguished from its meaning. That is, a word is an appearance, an internal or mental thing, whereas its meaning, at least in many cases, is an external 'real' entity. For instance, 'John' is a word whose meaning or reference is a person, John. This observation is central to most treatments of intentionality; J. S. Mill in particular made it the focal point of his treatment of the meaning of proper names.

I think the ARD also has some significance for the philosophical issues surrounding intentionality. Consider Dennett's 2-bitser. This is a vending machine that accepts quarters. Dennett argues -- correctly, I believe -- that the 2-bitser can be said to 'represent' or 'mean' a quarter by its internal state that results from accepting a quarter, only by virtue of an on-looker that so interprets the state. That is, the 2-bitser's intentionality is derived, not intrinsic. Dennett suggests that this is true of all intentionality, even ours. However, the 2-bitser is not an ARD device, and we are. I think that the ARD feature may lead us out of mere derived intentionality, to intrinsic intentionality. Of course, this will in part hinge on just what we take intentionality to be. But one thing seems promising at the outset: the ARD has to it a built-in kind of aboutness or directedness. The appearance is about the reality. That is, a device with an ARD capability will have internal representa-

tions distinguishing an appearance A from the supposed reality R behind it. Then A is internally 'directed' toward R. For instance, R can be the 'meaning' of A to the device.

The ARD requires both appearance and reality to be, in some sense, internally represented, for the reasoning system is to explicitly reason about both, and so needs distinct tokens for the two. Both the word 'John' and the person John are to be represented, so that the reasoner can say or think that the former names the latter. This already makes us different from the 2-bitser, for the vending machine has no way to regard its state as standing for anything else. That is why there is only a derived notion of representation for it: we can interpret its state as representing a quarter.

However, if the 2-bitser were equipped with a camera and suitable internal mechanisms, it could relate the state that results from accepting a supposed quarter with its visual data formed from the camera's pointing at whatever is being pushed into its slot. That is, it could treat its 'accept' state as a name for the visually parsed datum, much as we may think of the name 'John' as attached to what we see before us (John, except of course that this is mediated by our eyes and brains, just as in the case of the 2-bitser's camera and associated mechanisms). Of course, our brains are vastly more complex than anything we currently can build into a machine, but that is another matter.

Now, one problem (of many) that surfaces here is error. Dennett points out that the 2-bitser can be fooled (with respect to its derived intentionality) by using instead of a US-quarter a quarter-balboa (identical in all significant respects to a quarter, but not acceptable to the machines owners). That is, the fooling is really with respect to people, not the 2-bitser: it is too dumb to be fooled, since it has no intentions, no interpretations of its own to be gotten around. It is people's intentions and interpretations that are fooled.

The use of a camera may offset this, by reading the inscription on the quarter-balboa and rejecting it. This might proceed by comparing the 'accept' state and the visual data and deciding that the two don't match: the former says 'US-quarter' and the latter 'quarter-balboa'. This then could cause the machine to return the quarter-balboa with a stern vocalized message to the person

whoever inserted it. In colloquial terms, the 2-bitser would have caught its own mistake.

The matter will not rest there, however. For one thing, matching a canonical picture of a quarter is also not foolproof, it is no guarantee of being produced in the proper way by the U.S. Mint. Now of course, people are not very good at assessing this either. But at least we can understand the concept of being a 'real' (US) quarter, and recognize that this is different from our mere error-prone judgement that something is a quarter. Or so we tell ourselves. Dennett and others seem to think not, that this is an illusion about ourselves. And certainly it is difficult to say what it is that constitutes the 'real' meaning of terms, apart from our judgements.

What we would like are truth conditions for being a quarter, etc. If the conditions reside in our own judgements (verificationism) then how can we ever be wrong? And if not in our minds then where and what good do they do us? Fodor and others have struggled to make good on an internal notion of error, without apparent success. What we want is to be able to be wrong and to recognize this. But to recognize it is apparently to have in mind the right answer and contrast it with the wrong. Yet if we have the right answer how do we ever come to choose the wrong one in the first place?

The causal theory of reference tries to capture this by means of suitable generalizations based on the key causal features in the growth of a term's use, e.g., paradigmatic examples of things that came to be called quarters. This has several very difficult aspects, though it is perhaps the most robust theory around right now. One of the originators of the theory, Hilary Putnam, seems to have abandoned it in favor of the view that reference is never fully and finally fixed in an external reality but rather is always relative to a language user's point of view. This would appear to be the case for our camera-equipped 2-bitser, for instance. One advantage humans may have over this souped-up 2-bitser is that we can adapt our usage as we learn more; we may start with a rather simplistic notion of quarter and then over time come to employ a far more subtle notion. For this a recursive ARD capability seems just the thing, something I call 'reflection' on a 'presumed external thing'.

Be that as it may -- and I think the ARD approach has still more to offer on this -- I think it already is apparent that there is a significant behavioral watershed when a device is able to employ the ARD. There is an internal directedness that makes symbols symbolic to the device itself, and furthermore this has behavioral adaptiveness in that a finer range of distinctions and error corrections (as in returning the quarter-balboa) becomes possible.

(1) Dennett, D. [1987] Evolution, error, and intentionality. In D. Dennett, *The Intentional Stance*. MIT, 287-321.

(2) Devitt, M. and Sterelny, K. [1987] *Language and Reality*. MIT.

(3) Fodor, J. [1987] *Psychosemantics*. MIT.

(4) Perlis, D. [1987] How can a program mean? IJCAI-87.

(5) Putnam, H. [1988] *Representation and Reality*. MIT.